SUSMAN GODFREY L.L.P.

October 24, 2024

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York
New York, New York 10007

Re: *The New York Times Company v. Microsoft Corporation, et al.*,
Case No.: 23-cv-11195-SHS-OTW: Discovery Dispute Regarding OpenAI's Responses to The Times's Second and Third RFPs

Dear Magistrate Judge Wang:

Plaintiff The New York Times Company ("The Times") requests that a dispute over OpenAI's responses to document requests be heard at the upcoming October 30, 2024 discovery conference. OpenAI should be ordered to produce documents in response to four requests contained in The Times's Second and Third Sets of Requests for Production ("RFPs").[1] These four RFPs were briefed in the parties' agenda for the upcoming conference. Dkt. 262 at 28-35.

The parties have resolved many of the disputed RFPs listed in the agenda, but were unable to resolve the four RFPs addressed in this letter. While The Times remains open to resolving these disputes without court intervention, The Times wishes to provide additional background for the Court's consideration in advance of the conference.

1. **Documents Concerning OpenAI's Knowledge and Reliance on Websites that Copy Times Content (RFPs 84-85)**

OpenAI has so far refused to produce any documents in response to RFPs 84 and 85.

RFP 84 seeks documents "concerning Your knowledge of DNyuz (www.dnuyz.com), including its copying of Times Content and Your use of DNyuz content for any purpose, such as training, fine tuning, retrieval augmented generation, and any other refinement or development processes." DNyuz "is a notorious pirater of news" that has made "significant amounts of money by copy-pasting work by publications – including the NYT – and running ads on stolen content."[2] Public reports have explained how ChatGPT relies on and pulls from DNyuz when providing responses to user queries, including to copy Times content. *Id.* The Times reasonably seeks documents concerning OpenAI's knowledge of DNyuz and OpenAI's use of the site, which is relevant to The Times's allegations of willful infringement, among other issues. FAC ¶¶ 124-26.

[1] The parties met and conferred regarding these issues via videoconference on August 26, August 29, October 18, and October 24, and the parties exchanged correspondence as well. The Times's Second and Third RFPs are attached as Exhibits A and B, and OpenAI's Responses and Objections to these RFPs are attached as Exhibits C and D. Attached as Exhibit E is a July 31 letter from The Times. Attached as Exhibit F is an August 20 letter from OpenAI. Attached as Exhibit G are email exchanges between the parties.

[2] https://futurism.com/chatgpt-plagiarized-nyt-articles.

OpenAI does not dispute the relevance of this request. To the contrary, OpenAI told this Court in the parties' agenda that "OpenAI has already run 'dnyuz' as a search term and reviewed the resulting hits." Dkt. 262 at 34. Despite that representation, OpenAI has refused to provide any hit counts to The Times, and more recently, OpenAI implied that it may not have run this search term after all. *See* Ex. G at 8 (OpenAI suggesting The Times should now formally propose "dnyuz" as a search term in order to discover the hit counts). Setting aside OpenAI's potential about-face on search terms, OpenAI still refuses to confirm that it will produce documents in response to this RFP. *Id.*

RFP 85 is similar; it seeks "documents concerning Your knowledge of websites besides DNyuz that regularly copy publisher content without permission, and Your use of such websites for any purpose." OpenAI claims this RFP is overbroad because it seeks documents concerning an "unbounded universe of third-party websites." Ex. G at 11. But that argument mischaracterizes the request. Not every website is known to pirate publisher content. Moreover, responsive documents can easily be located by using search terms tied to the at-issue misconduct, such as terms that will capture documents relating to copying and plagiarism.

2. Documents Concerning OpenAI's Use of Web Crawlers to Access Times Content (RFP 26)

RFP 26 seeks documents "concerning Your use or knowledge of web crawlers, bots, spiders, user agents, and related tools to access The Times's Content, including Times websites and digital products." The Times through meet and confers have asked OpenAI to commit to producing documents concerning the following web crawlers: "GPTBOT, ChatGPT-User, OAI-SearchBot, CommonCrawl bot, and crawlers used by Microsoft or developers of CustomGPTs." Ex. G at 16. OpenAI has agreed to produce some documents in response to this request, but not the full scope of relevant documents.

The remaining dispute is whether OpenAI should produce documents regarding web crawlers used by Microsoft or by developers of CustomGPTs. It should. The Times alleges that "Microsoft actively gathers copies of the Times Works used to generate such results in the process of crawling the web to create the index for its Bing search engine." FAC ¶ 95. Relatedly, The Times alleges that "Microsoft and OpenAI continue to create unauthorized copies of Times Works in the form of synthetic search results returned by their Bing Chat and Browse with Bing products." *Id.* Documents produced by Microsoft reveal [redacted] Ex. H at -11, -14 (Agreement [redacted] OpenAI should search for and produce any documents it finds concerning its knowledge of and use of Microsoft web crawlers for purposes of accessing Times content.

OpenAI should also produce documents concerning web crawlers used by developers of customGPTs. As explained in the Daily News complaint, "OpenAI's Custom GPT Store contains numerous Custom GPTs specifically designed to circumvent the Publishers' paywalls", including a 'Remove Paywall' Custom GPT" and a "'News Summarizer' Custom GPT" that "encourages users to save on subscription costs and skip paywalls." Daily News Compl. ¶ 147. OpenAI can and should conduct an investigation into its knowledge of web crawlers used by CustomGPTs to access Times content, and produce any responsive documents it finds.

3. Documents Concerning any Removal of Copyright Management Information ("CMI") (RFP 18)

RFP 18 seeks "Documents concerning the removal of any content, including information regarding authorship, date of publication, publishing entity, title, copyright notices, and terms and conditions for use of works, from Training Datasets before, during, or after training of Defendants' Generative AI Model(s)." This request is relevant to The Times's DMCA claim, which includes claims for removing or altering CMI. *See* FAC ¶¶ 181-91. The parties have negotiated revised wording for this RFP, and OpenAI has agreed to produce documents concerning any removal of CMI in connection with the model training process. Ex. G at 4.

But one dispute remains. OpenAI refuses to produce documents concerning any removal of CMI in connection with retrieval augmented generation ("RAG"). *Id.* RAG enables OpenAI to integrate content from the "live" web—including The Times's website—with their large language models. By using RAG, ChatGPT can generate answers to queries about current events and other information that postdates the training of the models. FAC ¶¶ 81, 108-23, 163, 179.

OpenAI has not (and cannot) argue that the requested documents are irrelevant. The Times alleges that OpenAI's DMCA violations include the removal of CMI in connection with RAG. *See, e.g.*, FAC ¶ 185 (alleging that OpenAI removed CMI "when scraping Times Works from The Times's websites and generating copies or derivatives of Times Works as output for the Browse with Bing and Bing Chat offerings"). OpenAI has acknowledged the relevance of such documents, agreeing to produce documents concerning RAG in response to other RFPs, including RFPs 12 and 13. *See* Dkt. 147 at 3.

OpenAI instead takes the position that the language of this RFP excludes documents related to RAG. That interpretation is misguided because the RFP expressly seeks documents concerning the removal of CMI from Times Works "after" training, which includes RAG. In any event, OpenAI's argument regarding the wording of this RFP could be addressed by serving a new RFP, but that approach unnecessarily injects delay into discovery. The more efficient approach is for OpenAI to produce the requested documents, particularly when OpenAI cannot articulate a substantive objection to doing so.

Respectfully,

/s/ Ian B. Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

/s/ Steven Lieberman
Steven Lieberman
Rothwell, Figg, Ernst & Manbeck

cc: All Counsel of Record (via ECF)